nature biotechnology

Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube

Fan Zhang[1,7], Lena Christiansen[1,7], Jerushah Thomas[1], Dmitry Pokholok[1], Ros Jackson[2], Natalie Morrell[2], Yannan Zhao[3], Melissa Wiley[3], Emily Welch[3], Erich Jaeger[4], Ana Granat[4], Steven J Norberg[1], Aaron Halpern[4], Maria C Rogert[3], Mostafa Ronaghi[1], Jay Shendure[5], Niall Gormley[2], Kevin L Gunderson[6] & Frank J Steemers[1]

Haplotype-resolved genome sequencing promises to unlock a wealth of information in population and medical genetics. However, for the vast majority of genomes sequenced to date, haplotypes have not been determined because of cumbersome haplotyping workflows that require fractions of the genome to be sequenced in a large number of compartments. Here we demonstrate barcode partitioning of long DNA molecules in a single compartment using "on-bead" barcoded tagmentation. The key to the method that we call "contiguity preserving transposition" sequencing on beads (CPTv2-seq) is transposon-mediated transfer of homogenous populations of barcodes from beads to individual long DNA molecules that get fragmented at the same time (tagmentation). These are then processed to sequencing libraries wherein all sequencing reads originating from each long DNA molecule share a common barcode. Single-tube, bulk processing of long DNA molecules with ~150,000 different barcoded bead types provides a barcode-linked read structure that reveals long-range molecular contiguity. This technology provides a simple, rapid, plate-scalable and automatable route to accurate, haplotype-resolved sequencing, and phasing of structural variants of the genome.

Standard sequencing library preparations typically fragment a genome into several-hundred-base-pair-long library inserts, effectively removing long-range (10–500 kilobases (kb)) contiguity information[1]. However, the importance of long-range information is evident in applications that require haplotype-resolved genome sequencing (phasing), assembly, and structural variant detection[2]. For example, inheritance patterns of genetic variation in complex traits are often influenced by interactions among multiple genes and alleles across long distances. Examination of phased variants, as well as their *cis*- or *trans*-interactions, are critical for a greater understanding of the genetic basis of complex phenotypes[2–6]. Moreover, resolution of long-range information at the individual molecular level within complex samples, such as cancer samples, is essential to assemble and phase variants of subpopulations of cells. Structural variant and gene-fusion detection are also facilitated by long-range contiguity information[7,8]. Such rearrangements can be genetic drivers and important diagnostic biomarkers in cancers and other diseases[9,10]. Finally, phase information is needed to discern compound heterozygosity in thousands of recessive Mendelian diseases.

A range of experimental and computational[2,11] methods have been developed to capture long-range information in the genome, including dilution haplotyping[12–14], paired-end, mate-pair[15–17], synthetic long read sequencing[18–21], and Hi-C[22]. Methods are often used in combination with one another to improve assemblies, as different length-scales or read structures provide complementary information[8,23,24]. The principle of dilution haplotyping is based on the separation of parental copies into many different physical compartments, such that each compartment contains a sub-haploid genome equivalent[25]. When combined with compartment-specific barcoding and library preparation, long-range information can easily be inferred from library elements sharing the same compartmental barcode[26–28]. However, a major shortcoming of these approaches is the need for substantial automation or complicated microfluidics when scaling up the number of barcoded compartments or genomes. Other technologies, such as single-molecule long-read technologies (e.g., SMRT-sequencing and nanopore sequencing) can provide long-range information[12,21,29,30], but sequence read accuracy, read length, and protocol complexity are currently barriers for its widespread adoption. Thus, there is a need for improved methods that can deliver fully phased genomes, with accurate sequence as well as structural variant detection. It is essential that such methods be simple, scalable, complete, and accurate, if they are to be broadly adopted in the context of population-scale human genome sequencing projects.

Here, we demonstrate that haplotype-resolved sequencing can be performed in one physical compartment. Our method performs many individually barcoded library preparations from long DNA molecules simultaneously, within a single-tube reaction. Consequently, it retains plate-based scalability with automation integration, a desirable feature for processing large numbers of samples. We demonstrate that this method results in accurate phasing of variants with only nanograms of

[1]Advanced Research Department, Illumina, San Diego, California, USA. [2]Technology Development Department, Illumina, Little Chesterford, Essex, UK. [3]Technology Development, Illumina, San Diego, California, USA. [4]Gene Expression Department, Illumina, San Francisco, California, USA. [5]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [6]Encodia, Inc., San Diego, California, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to F.J.S. (fsteemers@illumina.com).

Received 14 October 2016; accepted 10 May 2017; published online 26 June 2017; doi:10.1038/nbt.3897

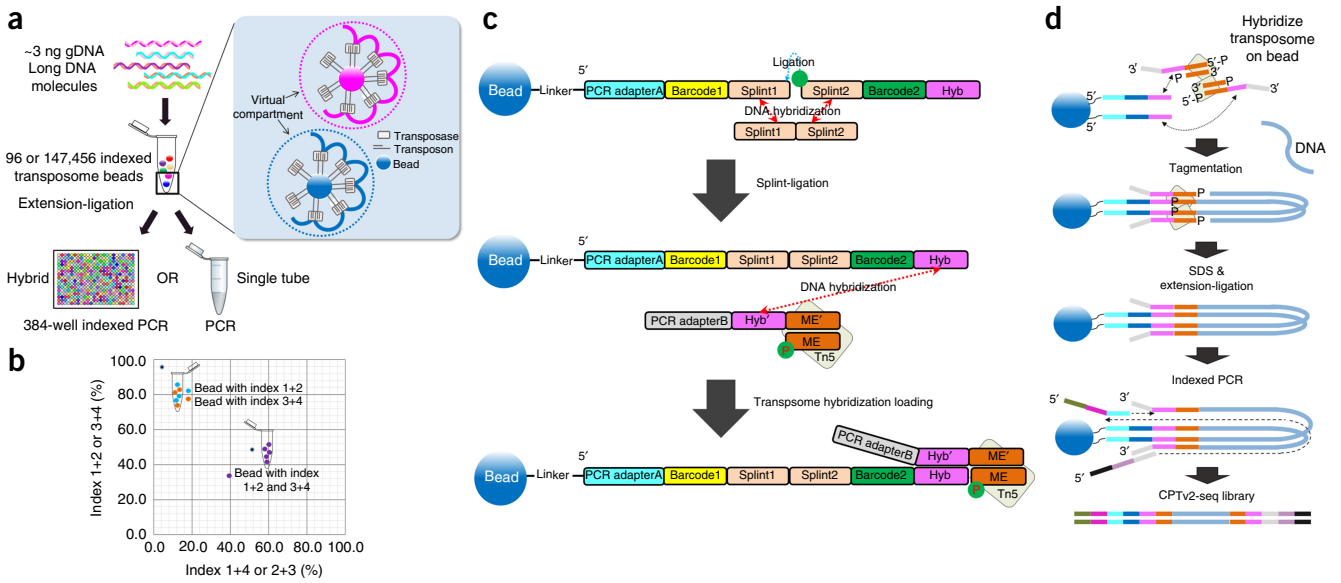

Figure 1 Summary of the bead-based indexing workflow and the intra- vs. inter-bead tagmentation test. (**a**) Overview of on-beads tagmentation and indexing. Tagmentation on beads in a single tube is followed by PCR (single-tube method), or a 384-well barcoded PCR (hybrid method). Close-up visualization shows clonally barcoded bead transposition with 'virtual compartments' inside a single physical compartment. The hybrid approach starts with the preparation of 96 uniquely indexed, biotinylated transposomes that were individually loaded on streptavidin beads and combined to make a bead pool used for transposition. After transposition, the beads were distributed into 384 wells for a second round of indexing by PCR. In the single-tube method, the bead pool with ~150K distinct barcodes was synthesized using a split-and-pool strategy. Transposomes loaded with a universal transposon were hybridized via its common hybrid′ (Hyb′) sequence to the hybrid (Hyb) sequence attached to the barcoded beads. A long DNA molecule in proximity to a single bead is tagmented multiple times by transposomes on that bead, generating individual libraries with the same barcode. (**b**) Results of initial feasibility experiment with two bead types, barcode 1+2 and barcode 3+4 demonstrating that target DNA is clonally barcoded through transposition on a single bead. (**c**) Combinatorial transposome bead pool preparation used for single-tube approach (for simplicity, only one bead type and a half of the dimer transposome complex are illustrated). A split-and-pool synthesis strategy was used to make a ~150K (384 × 384) bead pool, with each bead having a unique index combination. Initially, 384 uniquely indexed beads were individually prepared with barcode 1 sequences (yellow box). The beads were pooled and split into a new 384-well plate for splint-assisted ligation (red-dash arrows) with a set of 5′ phosphorylated oligos, containing a unique barcode 2 sequence (green box). The transposomes, formed with a universal transposon, were hybridized to a common hyb sequence (purple box) on the beads, generating the 150K barcoded transposition-active bead pool. (**d**) Workflow for library generation. The 150K barcoded transposition-active bead pool was prepared as described above (only halves of transposome dimers are illustrated). Long DNA molecules were mixed with the bead pool and tagmented by transposomes on the bead surface. After transposition, Tn5 transposase was released with SDS facilitating the extension/ligation steps to fill the gaps (9-bp gap from transposition) and ligating the bead index oligo to the library. Finally, PCR introduced the sequencing adapters.

DNA input, covering >99% of all heterozygous single-nucleotide polymorphisms (SNPs) genome-wide with megabase-scale phasing blocks. Additionally, we show the utility of barcode-linked reads in accurately detecting long-range structural variants with parental specificity.

We recently introduced the concept of CPT-seq, in which contiguity-preserving Tn5 transposition is combined with combinatorial indexing to obtain long-range information[26]. Here we investigate whether bead-immobilized barcodes can be transferred to a single, long DNA molecule using transposition, effectively barcoding the sequence content of that single DNA molecule, or a limited number of DNA molecules, that happen to interact with the same bead. Highly parallel labeling of many long DNA molecules was achieved by including a large number of distinct barcoded beads in a single tube, wherein each bead type holds many copies of the same barcode (**Fig. 1a**). Thus, mapping of short reads that contain an identical barcode sequence should produce clusters ('islands') that map in close proximity on a reference genome, with each cluster derived from the same long DNA molecule.

To test this idea, we prepared two different populations of barcoded transposome-functionalized beads (bearing either barcodes 1 and 2 or barcodes 3 and 4). The two bead types were pooled, mixed with DNA, and tagmentation was initiated. During tagmentation, the barcoded-immobilized transposon is transferred to the target DNA. If the entirety of each single DNA molecule interacts with only a single bead, the majority of sequenced fragments will be flanked by barcode 1/2 or barcode 3/4 combinations since tagmentation is intra-bead (**Fig. 1b** and **Supplementary Table 1**). Alternatively, if a single DNA molecule interacts and gets tagmented with multiple beads, the resultant sequenced fragments will, in part, be flanked by inter-bead barcodes (barcode 1/3 and 1/4 or barcode 2/3 and 2/4). As a negative control, we pre-mixed all four transposon-barcodes before transposome formation and then immobilized them on beads, essentially scrambling all possible combinations on each bead (**Fig. 1b** and **Supplementary Table 1**). We observed that the majority of the reads (>95%) were indexed through barcode transfer from single beads. As expected, the negative control gave all possible barcode combinations at approximately equal rates. These results suggest that barcode-linked reads can be generated by exposing a DNA sample to a complex population of barcoded beads within a single tube.

To apply bead-based barcode partitioning to phasing, we designed two workflows (**Fig. 1a** and **Supplementary Table 2**). Initially, we implemented a 'hybrid' approach in which a single tube of 96 different barcoded bead types was used to tagment a DNA sample. Analogous to our CPT-seq technology[26,31], the first barcode was incorporated into the genomic DNA library by transposition, and the second barcode was incorporated during barcoded PCR in 384-well plate(s).

Table 1 Summary of phasing results for the hybrid and the one-tube CPT-seq on beads methods

DNA source	NA12878 subsample	NA12878		
DNA prep	Gentra	Gentra	Freshly prepared DNA	Freshly prepared DNA
Approach	Hybrid	Hybrid	Hybrid	One tube
Barcodes	73,728	110,592	73,728	147,456
Read length	2 × 76	2 × 76	2 × 76	2 × 76
Number of reads (millions)	623	934	679	648
Mapped bases (Gb)/mapped %	75/80%	112/80%	86/84%	73/75%
Uniqueness %/duplicates %	78/22	62/38	68/32	79/21
Mean depth of coverage (duplicates removed)	19.5	23.1	19.5	19.2
Mean DNA/partition (Mb)	7	10	15	6
Informative linked reads[a] N50 (kb)	59.7	63.2	97.3	58.5
Mean number of linked reads	8	11	11	5
N50/max. of linked read region (kb)	42.9/276	47.7/282	73.5/748	34.9/339
hetSNPs phased (%)	98.4	99.3	99.5	98
Phasing block N50 (Mb)	0.92	1.19	2.43	1.14
Longest phasing block (Mb)	2.52	3.28	5.48	3.46
Short switch error rate (%)	0.086	0.045	0.039	0.13
Long switch error rate (%)	0.018	0.0056	0.0014	0.0085

[a]The linked reads cover at least two heterozygous SNPs.

In this manner, a set of virtual partitions, equal in number to the barcoded transposition reactions in the first step, can be defined within each physical compartment. Sequencing the same set of barcodes from each 384-well plate on individual lanes of a HiSeq sequencing system provided an additional level of barcode scalability as pseudo-barcodes were assigned per lane, which provided 36,864 (96 × 384 × 1 lane), 73,728 (96 × 384 × 2 lanes), and 110,592 (96 × 384 × 3 lanes) index partition data sets. After PCR, sequencing confirmed the successful generation of the expected barcoded libraries (**Supplementary Fig. 1a**). We observed consistent library insert sizes of ~300 bp (**Supplementary Fig. 2**); the insert size could be modulated through the transposome density on the bead (data not shown). The transposome beads rapidly (<10 min) bound over 95% of the initial DNA input, and tagmented DNA remained attached to the beads. We extensively optimized the assay for DNA input, number of beads, PCR cycles, and various methods to reduce losses and improve robustness. In general, we observed that larger tagmentation volumes, gentle handling of DNA, a higher number of barcodes or partitions, and a lower number of DNA molecules per bead, all resulted in longer islands and higher phasing quality (data not shown).

As a proof-of-concept test, we applied our CPT-seq-on-beads assay to HapMap sample NA12878 for which gold-standard haplotype information is available. Three nanograms of input DNA were processed through the described hybrid method and sequenced on a HiSeq 4000 using a 2 × 76 paired-end reads workflow. Data were analyzed as previously described[26] and are reported at different coverage levels in **Table 1**. The first indication of the efficacy of barcode-linked reads was observed in the island-like distribution pattern of mapped reads for each barcode sequence (**Supplementary Fig. 3**). Additionally, a plot of the distances between adjacent alignments for a given barcode sequence exhibited a bimodal distribution, in which the distances from within an island (proximal reads) contributed to the peak centered around 2 kb, and the distance between islands or sporadic reads (distal reads) contributed to the second peak centered around 4 Mb (**Supplementary Fig. 4**). A high fraction of proximal versus distal reads indicates that the majority of reads (>90%) were in an island structure.

The N50 of informative islands, containing more than one heterozygous SNP, was 63 kb (N50 is the largest value for which items of that length or longer make up at least 50% of the sum of all items' lengths.). About 10 Mb of the genome, or 0.3% haploid equivalents, was distributed in each barcoded partition. The assignment of the haplotypes for the islands was performed by H-BOP[32] (**Supplementary Fig. 5**). The N50 of the assembled haplotype blocks was in the megabase-scale, showing high accuracy with short and long switch error rates of 0.045% and 0.0056%, respectively. The heterozygous SNP (hetSNP) coverage was >99% without any imputation from 20× genome coverage (**Table 1**). We calculated phasing yield and accuracy as a function of SNP distance[26] (**Supplementary Fig. 6**). Phasing yield, the probability that heterozygous SNP pairs are on the same phasing block as a function of the distance between them, showed that 99% of the SNPs were in correct phase at distances up to 80 kb. Phasing accuracy, the probability that a SNP pair is phased correctly as a function of distance, was 99.85% at 80 kb. Phasing at only 10× genome coverage covered >95% hetSNPs with slightly lower accuracy (~0.16% short switch errors; data not shown).

To improve the quality of the DNA, we prepared DNA directly[33] before the experiment, generating a substantially greater fraction of longer islands (300–500 kb) compared to older DNA stocks purified with the Gentra DNA isolation kit (**Supplementary Fig. 7**). At 20× genome coverage, 99.5% of the hetSNPs were covered with long switch error rates down to 0.0014, or ~1 error in 100 Mb. The phasing block N50 was 2.43 Mb and the longest phasing block was 5.48 Mb.

In the hybrid method, each barcoded transposome bead was prepared individually and then pooled (as there were only 96 different transposons). To scale to ~150,000 bead barcodes for the single-tube implementation, we developed a split-and-pool combinatorial synthesis approach to prepare 147,456 (384 × 384) barcoded bead partitions (**Fig. 1c**). We then developed a method to convert a complex combinatorial bead pool, to an enzyme-active barcoded transposome bead pool in a single step.

With the single-tube method using the combinatorial transposome bead pool (**Fig. 1d**), we generated a library from the NA12878 sample (**Supplementary Fig. 1b**) with a relatively uniform distribution of the read numbers (kurtosis = 1.02, skewness = 0.64) across the ~150,000 barcode space (**Supplementary Fig. 8**). At similar 20× genome coverage, using freshly prepared DNA, 98% hetSNPs were phased with short and long switch error rates of 0.13% and 0.0085%, respectively, similar to the hybrid approach (**Table 1**).

We next explored the capability of barcode-linked reads to identify structural variants of NA12878 using the hybrid approach on Gentra-prepared DNA. We expected the number of barcodes containing reads within the deletion loci to be lower compared to within adjacent genomic regions. Because unique barcodes were assigned to

the respective haplotype using the linked read information, the heterozygous deletions should show significantly unbalanced numbers of barcodes for the two haplotypes. Overall 296 regions >1 kb showed significant signals for possible heterozygous deletions (*P*-value of barcode counting < 0.001, and *P*-value of heterozygous deletion from binomial test <0.01). The data set was also analyzed by Manta[34], which

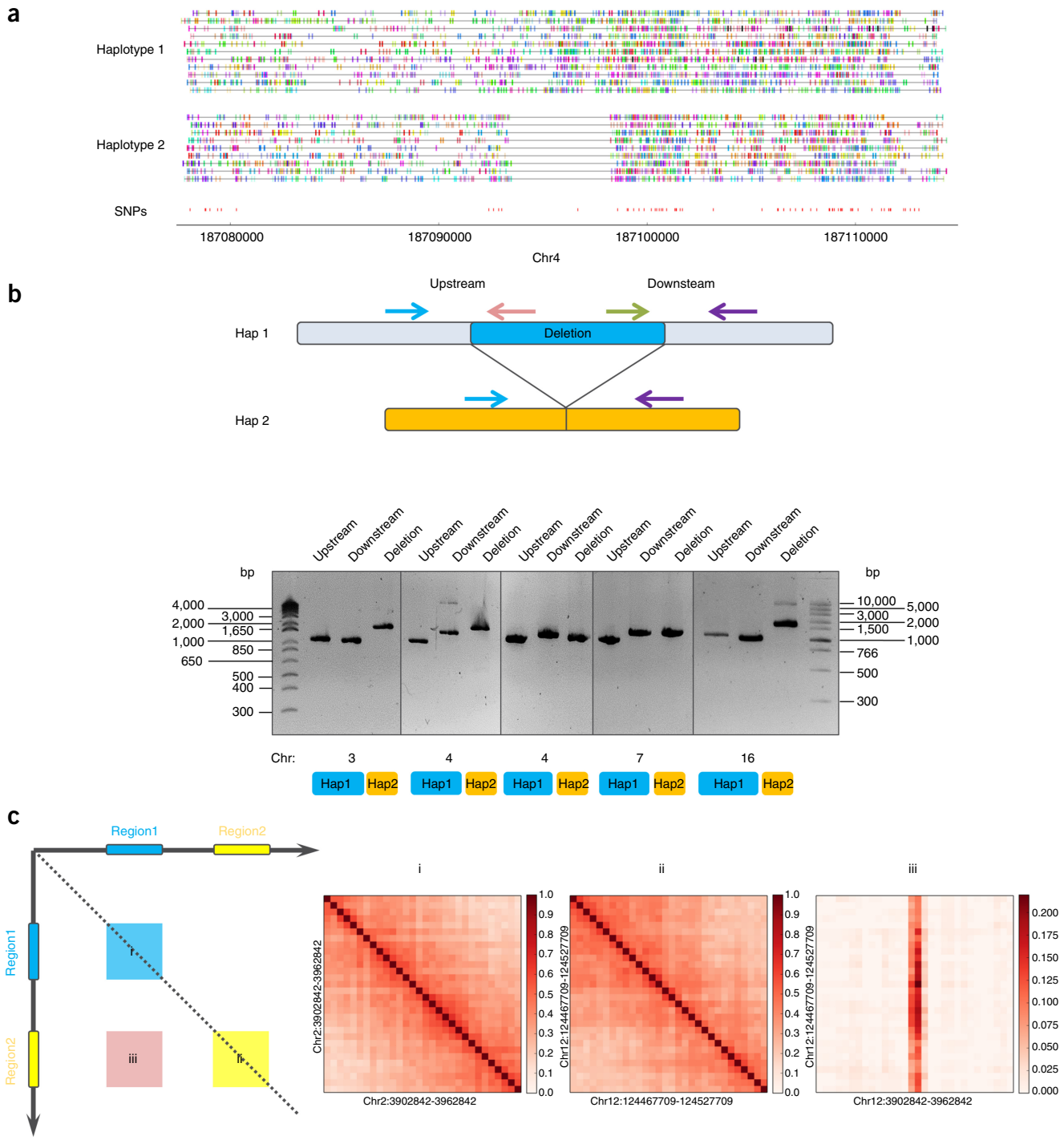

Figure 2 Detection of deletion and interchromosomal translocation using linked read information. (**a**) Example of a ~4-kbp heterozygous deletion on chromosome 4 of NA12878. Every sequencing read is shown as a small vertical bar with each distinct color assigned to a specific barcode. In haplotype 1 the linked reads scatter across the whole region, and in haplotype 2 there is a gap in linked read coverage confirming the heterozygous deletion. (**b**) PCR confirmation of five randomly selected heterozygous deletion alleles discovered by linked read analysis. Specific PCR primer pairs were designed upstream and downstream of each heterozygous deletion. For an allele without a deletion (Hap1), PCR amplified two fragments of the genome of ~1 kb in size for all regions. For each deletion allele (Hap2), outside primer pairs (blue and purple primers) generated a 1- to 2-kb product band, confirming the fusion of two regions of the genome. (**c**) Example of an interchromosomal translocation event between chromosome 2 and chromosome 12 in the HeLa cell genome. Left panel illustrates a scheme of the Jaccard index of barcode sharing scanned across the translocation regions. Panels i and ii, the heatmap of barcode sharing from self-crossing of region 1 (Chr2: 3902842-3962842) and region 2 (Chr12: 124467709-1245527709), respectively; iii, a heatmap of barcode sharing between region1 and region2, showing a strong signal as a red band in the middle of the heatmap.

Table 2 Examples of the heterozygous deletions detected for NA12878 using barcode analysis and Manta[34]

Chromosome	Region	PCR confirmation	Manta detection
3	65189000-65213999	Yes	65188869-65214748
4	116167000-116176999	Yes	NA
4	187094000-187097999	Yes	NA
7	110182000-110187999	Yes	110181965-110188432
16	62545000-62549999	Yes	62544333-62550668

discovered 570 deletions >500 bp without leveraging linked reads information. There were 107 shared regions between these two analyses. An example of a ~4-kbp heterozygous deletion on chromosome 4 is displayed in **Figure 2a**, with the reads separated into different haplotypes according to their respective barcodes. Five randomly selected deletions newly discovered by CPTv2-seq were confirmed by PCR (**Supplementary Table 2**) and visual inspection (**Table 2**, **Fig. 2b** and **Supplementary Fig. 9**), as well as by the sequencing depth analysis (**Supplementary Fig. 10**) using the trio data set (NA12878, NA12891, NA12892) from the platinum genomes project[35]. 215 of 296 observed heterozygous deletions were confirmed by single-molecule sequencing technology[30].

More complex structural variants were detected by using a metric of statistically significant barcode sharing between regions of the genome. The barcode-linked libraries, or islands, retain proximity information and can be used to detect structural variants[8] and aid assembly[36]. In general, the closer the regions are arranged in the genome, the higher the probability they share the same barcodes with each other. The power of barcode sharing for interchromosomal translocation detection was evaluated by sequencing the genome of the well-characterized HeLa cell line[37] using the hybrid approach. The mapped reads were scanned using a 2-kb sliding window along the genome, and a Jaccard index was calculated for the barcodes shared between the window pairs. The number of barcodes shared between two regions provided information about their proximity, as reads in proximity to one another on the genome scale were more likely to share the same barcode. A higher Jaccard index was observed for a known interchromosomal translocation previously reported[37] (**Fig. 2c**), and 17 out of a total of 20 interchromosomal translocations could be reproduced with the current approach (**Supplementary Fig. 11**).

The single-tube CPT-seq method described here provides haplotype-resolved sequencing information with a simple protocol. We show that parental DNA molecules can be differentiated in a single physical compartment using bead-based barcode partitioning, in which intra-bead molecular barcoding is highly favored over inter-bead barcoding. This feature enables high-throughput plate-based automation (96–1,536 samples) with short assay times (<3 h and 30 min hands-on). Compared to dilution haplotyping, our approach shifts the challenge from making many physical partitions to simply building a high-complexity barcoded-bead pool. Manufacturing such a bead pool is very similar to making a complex bead pool for Illumina's Infinium bead array products[38] or using combinatorial split-and-pool synthesis techniques as demonstrated before[39].

Our bead-based CPT-seq method yields raw linked reads with >99% haplotyping accuracy (**Supplementary Fig. 12**), and generates megabase phasing blocks with low switch error rates, and are comparable to recently reported methods[8]. The genomic DNA input amount of 2–5 ng and around 24 million beads is a good balance between effective separation of paternal and maternal DNA of the same allelic region and generating sufficient library material for sequencing. Several parameters can be further optimized to improve phasing performance: the island read structure (coverage and length of islands and unique reads) and the bead pool geometry and complexity. The relatively low uniqueness of sequencing data (~64–80%) was likely due to the low DNA input (~3 ng) combined with library amplification and clustering bias. There was also a theoretical inverse relationship between the coverage of the islands and the unique reads percentile. In general, we observed higher coverage and longer islands with lower uniqueness. This was confirmed with data sets with ~40% uniqueness (data not shown), in which the fraction of long islands between 300 and 500 kb increased even further resulting in N50s up to 8 Mb with >95% SNP coverage. This exemplifies the point that longer islands can substantially increase the size of the assembled haplotyping blocks. Finally, the variant calling from the barcode-linked library was obviously compromised with such low uniqueness. We report comparable coverage uniformity and SNP/INDEL recall and precision as in recently published whole genome haplotyping data by Zheng *et al.*[8] (**Supplementary Fig. 13** and **Supplementary Table 3**). Not unexpectedly, the uniformity and structural variant calling for these two sequencing methods were lower compared to the non-amplification-based TruSeq PCR-free workflow (**Supplementary Table 3**). Alignment of short reads has limitations in low complexity and repeat regions and therefore these genomic regions were not covered well. Notably, our method produced a lower number of chimeras, sequencing errors, and regions that were extensively overamplified compared to random-primer-based methods. SNP recall and precision rate of our 27× coverage data set from the hybrid approach compared to the platinum reference genome[35] was 94.1% and 99.6%, and for indels 82% and 93.1%, respectively, with slightly better metrics for the single-tube approach (**Supplementary Table 3**). Improvements in fraction uniqueness and read length will enhance the recall and precision metrics.

Library construction and sequencing is a random subsampling process with inevitable fragment loss[40]. In the single-tube approach, the capture efficiency of DNA with the transposomes immobilized on beads could be further optimized to improve the tagmentation, which is expected to generate more fragments within the island. One possible solution to improve coverage is to generate redundancy for each tagmented fragment. For example, low-bias amplification methods, such as *in vitro* transcription[41], and common and random primer amplification can generate shotgun reads over the individual fragments of the island, thereby compensating for loss from the subsampling process and improving unique reads. This shotgun amplification approach, with an engineered Tn5 showing less transposition bias[42], are currently being explored in order to improve variant calling and phasing. The same approach could potentially improve the sequencing quality of read 2 of the paired-end sequencing (shown as the gDNA on the right side of **Supplementary Fig. 1**), which is currently the largest source of unmapped reads. During read 2 preparation, the sequencing primer 2 competes with the internal library hairpin structure. Improvements to the sequencing recipe or the shotgun amplification approach effectively removes the hairpin potentially reducing the fraction of unmapped reads.

Captured contiguity of the long DNA molecules is determined by the quality of DNA in the assay and the diameter of the bead. The circumference of the current 3 µm bead is around 10 µm, allowing the contiguous capture of several 100 kb of DNA. Larger beads can improve the capture contiguity, but this has to be balanced with keeping the beads in suspension to avoid undesired cross-bead tagmentation. Future implementations can include >1 M bead partitions for improved molecular and parental resolution/accuracy with lower DNA content per partition.

We present an accurate single-tube phasing platform enabling plate-based, high-throughput, factory workflows. The method also provides structural organization of the genome using barcoded linked reads, as demonstrated with the detection of heterozygous deletions and interchromosomal translocations. One of the unique features of this method is that it is bead-based and can be potentially integrated with bisulfite sequencing to reveal haplotype-resolved methylation patterns. These technologies provide a simple, rapid, scalable, and highly automatable route toward routine, accurate haplotype-resolved sequencing, and assembly of the genome.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank the research, development, and software engineering departments at Illumina for sequencing technology development. Specifically, we would like to thank C.L. Pan at Illumina for the sequencing technology development and G. Bean, J. Leng, and S. Swamy at Illumina for data analysis software development. We would like to thank R. Daza for providing HeLa and NA12878 Gentra DNA preparations. The genome sequence described in this paper was derived from a HeLa cell line. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research.

AUTHOR CONTRIBUTIONS

F.J.S., K.L.G., and N.G. conceived the study; F.J.S. and M.C.R. oversaw the technology development. F.Z. and D.P. led the assay development, L.C., J.T., S.J.N., and D.P. performed the experiments, and analyzed the data. F.Z. performed the phasing analysis and wrote custom analysis software. A.G., E.J., R.J., and N.M. helped with the assay development. Y.Z., M.W., and E.W. prepared the bead pool. A.H. developed the data analysis pipeline. M.R. led the project coordination. F.J.S., F.Z., L.C., K.L.G., D.P., and J.S. co-wrote the paper. All authors contributed to the revision and review of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
2. Snyder, M.W., Adey, A., Kitzman, J.O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* **16**, 344–358 (2015).
3. Bansal, V., Tewhey, R., Topol, E.J. & Schork, N.J. The next phase in human genetics. *Nat. Biotechnol.* **29**, 38–39 (2011).
4. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
5. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
6. Nalls, M.A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
7. Peters, B.A. *et al.* Detection and phasing of single base de novo mutations in biopsies from human in vitro fertilized embryos by advanced whole-genome sequencing. *Genome Res.* **25**, 426–434 (2015).
8. Zheng, G.X.Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
9. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
10. Moncunill, V. *et al.* Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32**, 1106–1112 (2014).
11. Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6** (Suppl. 1), S13–S20 (2009).
12. Fan, H.C., Wang, J., Potanina, A. & Quake, S.R. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29**, 51–57 (2011).
13. Kaper, F. *et al.* Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc. Natl. Acad. Sci. USA* **110**, 5552–5557 (2013).
14. Kitzman, J.O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
15. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
16. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
17. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
18. Putnam, N.H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
19. Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.* **32**, 261–266 (2014).
20. Cao, H. *et al.* De novo assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* **33**, 617–622 (2015).
21. Peters, B.A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
22. Burton, J.N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
23. Sović, I., Križanović, K., Skala, K. & Šikić, M. Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. *Bioinformatics* **32**, 2582–2589 (2016).
24. Goodwin, S. *et al.* Oxford Nanopore Sequencing and de novo assembly of a eukaryotic genome. Preprint at *bioRxiv* https://doi.org/10.1101/013490 (2015).
25. Dear, P.H. & Cook, P.R. Happy mapping: a proposal for linkage mapping the human genome. *Nucleic Acids Res.* **17**, 6795–6807 (1989).
26. Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
27. Peters, B.A., Liu, J. & Drmanac, R. Co-barcoded sequence reads from long DNA fragments: a cost-effective solution for "perfect genome" sequencing. *Front. Genet.* **5**, 466 (2015).
28. Lan, F., Haliburton, J.R., Yuan, A. & Abate, A.R. Droplet barcoding for massively parallel single-molecule deep sequencing. *Nat. Commun.* **7**, 11784 (2016).
29. Loman, N.J., Quick, J. & Simpson, J.T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
30. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
31. Cusanovich, D.A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
32. Xie, M., Wang, J. & Jiang, T. A fast and accurate algorithm for single individual haplotyping. *BMC Syst. Biol.* **6** (Suppl. 2), S8 (2012).
33. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
34. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
35. Eberle, M.A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27**, 157–164 (2017).
36. Adey, A. *et al.* In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).
37. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).
38. Steemers, F.J. *et al.* Whole-genome genotyping with the single-base extension assay. *Nat. Methods* **3**, 31–33 (2006).
39. Furka, A., Sebestyén, F., Asgedom, M. & Dibó, G. General method for rapid synthesis of multicomponent peptide mixtures. *Int. J. Pept. Protein Res.* **37**, 487–493 (1991).
40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
41. Sos, B.C. *et al.* Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol.* **17**, 20 (2016).
42. Ason, B. & Reznikoff, W.S. DNA sequence bias during Tn5 transposition. *J. Mol. Biol.* **335**, 1213–1225 (2004).

ONLINE METHODS

Intra- versus inter-molecular bead transposition. Four individual transposons were designed with barcode 1 and barcode 3 at the P5 primer end, and barcode 2 and barcode 4 at the P7 end. Two bead types were prepared by making biotinylated transposomes with barcode 1+2 and barcode 3+4, and the resulting biotinylated products were bound to streptavidin beads, respectively. For the control all four transposomes were mixed and loaded on the beads. The sequencing library was generated by mixing DNA with beads as described in transposition and PCR sections below, followed by dual index sequencing on a MiSeq. After demultiplexing, the number of reads for different barcode combinations (1+4, 2+3, 1+2 and 3+4) were compared. The fragments generated by intra-bead transposition have the combination of either barcode1 and 2 or barcode 3 and 4, and the fragments from inter-bead transposition have the combination of barcode 1 and 4 or barcode 2 and 3 (**Supplementary Table 1**).

Assembly of 96 individually barcoded transposomes (hybrid method). The 96 barcoded transposons were formed by aliquoting eight individual barcoded oligonucleotides B16BS (IDT, HPLC purified, **Supplementary Table 2**) into rows A-H and 96 barcoded A18′ oligonucleotides (IDT, standard desalting, **Supplementary Table 2**) into each well, respectively, then annealing. Each B16BS transposon oligo contains a dual biotin at the 5′-end and the Tn5 Mosaic End (ME) sequence at the 3′-end, as well as adaptor sequences to make them compatible with the P7 Illumina sequencing end. Each A18 transposon oligo contains the complementary 19-bp ME sequence at the 5′ phosphorylated end, and have the adaptor sequence for the P5 side. 5 µl of 10 µM dual biotin B16BS oligonucleotide was mixed and annealed with 7 µl of 10 µM A18′ oligonucleotide. 2.5 µl of each the 96 unique annealed transposons were individually mixed on ice with 1 µl of 25 µM EZ-Tn5 transposase (Epicentre) and 21.5 µl of standard storage buffer (Epicentre) in a total volume of 25 µl and incubated at 37 °C overnight. Quality of complexes was assessed on an 8% TBE gel (Invitrogen) for 1 h at 100 V. A working stock of transposome complexes was prepared by diluting the transposomes to 100 nM in standard storage buffer and stored at −20 °C.

Preparation of 96 pooled and individually barcoded transposomes on beads (hybrid method). 300 µl (10 mg/ml) of magnetic streptavidin-coated M280 Dynabeads (Life Technologies) were washed one time with 1 ml of wash buffer (100 mM Tris-HCl; pH 7.5, 100 mM NaCl, 0.1% Tween 20). Beads were resuspended in 300 µl wash buffer and 3 µl of the resuspended beads was aliquotted into each well of a 96-well PCR plate containing 7.5 µl of wash buffer. 4.5 µl of each barcoded transposome (100 nM) was added to its corresponding well of the 96-well PCR plate and incubated at room temperature for 30 m on a Roto-Torque Heavy Duty Rotator (Cole-Parmer) on low speed. Beads were magnetized with a 96-well magnetic plate (Alpaqua) and washed two times with 200 µl wash buffer and then resuspended in 6 µl of working buffer (50 mM Tris-HCl; pH 7.5, 30 mM NaCl, 0.1% Tween 20). The full reaction was pooled by inverting the 96-well PCR plate into a collection plate (Illumina) and centrifuging 10 s at 250*g*. The pool was mixed by gentle swirling and gentle pipetting with a wide orifice tip (Mettler Toledo). 80 µl of the pool was taken with a wide orifice pipet tip for one transposition reaction.

Transposition of high molecule weight (HMW) and freshly prepared gDNA (hybrid method). All temperature incubations were performed in a PCR machine (Bio-Rad) unless otherwise stated. With a wide-orifice pipet tip (Mettler Toledo), 64 µl of working buffer containing 3 ng of HMW NA12878 gDNA (Gentra or freshly prepared from GM12878 cells ordered from Coriell Institute for Medical Research (Cat. No. GM12878)) or 3 ng of HMW Hela cell gDNA (Gentra prepared from ATCC Hela S3 cell (Cat. No. ATCC CCL-2.2)) was added to the beads and mixed by gentle flicking and inverting of the strip tube. The gDNA and beads were incubated at room temperature for 10 m at low speed on a rotator followed by the addition of 40 µl of 5× Tagment DNA buffer (TD, Epicentre). The sample was mixed by gentle flicking and inversion of strip tubes and incubated at a constant 55 °C for 10 m. After transposition, 40 µl of 5% SDS (Sigma Aldrich) in working buffer was added to the sample and mixed by gentle flicking and rotating, and the reaction incubated at 37 °C for 10 m. After SDS treatment, the strip tube was placed on a 96-well magnetic separator, and the supernatant was discarded. The beads were washed two times with 300 µl of wash buffer. The beads were resuspended in 50 µl of gap-fill ligation reagent ELM3 (Illumina) supplemented with 100 ng/µl BSA (NEB) and the reaction was incubated at 30 °C for 3 m followed by 16 °C for 30 m. After gap-fill ligation, the beads were washed two times with 300 µl wash buffer and resuspended in 25 µl of 0.4× TE containing 40 ng/µl BSA.

PCR barcoding (hybrid method). 2.5 µl of 4 µM Nextera i5 and i7 PCR primers with barcodes (IDT, standard desalting, **Supplementary Table 2**) was added to its respective well of a 384-well PCR plate on ice. 4 ml of a Phusion PCR master mix (ThermoFisher) was made for 15 µl total volume per reaction for a 384-well PCR plate. 3.33 µl of resuspended on-bead post-gap-fill ligated library was added to the 4 ml of master mix, and 10 µl was added to each well of a 384-well PCR plate (each well containing ~0.01 µl library). The following PCR parameters were used: an initial denaturation at 98 °C for 30 s, 9 cycles of denaturation at 98 °C for 10 s, annealing at 63 °C for 30 s, and extension at 72 °C for 30 s.

Post PCR purification and size selection (hybrid method). The 384-well PCR plate was pooled and purified with Zymo Clean & Concentrator-100 (Zymo Research). The following procedure was used for purification: 15 µl from 384 PCR reactions were pooled into a plastic solution basin and combined with 28 ml Zymo Binding Buffer. Half of the mixture was added to a 100 µg Zymo-Spin column and was centrifuged at 500*g* for 5 min; the remaining mixture volume was added to the column and centrifuged the same way. The column was washed two times with 2 ml Zymo Wash Buffer and eluted in 150 µl Zymo Elution Buffer. The entire 150 µl of concentrated and purified sample was added to 750 µl Zymo *Select-a-Size* binding buffer and transferred to an IC-S Zymo *Select-a-Size* column (Zymo Research). The column was centrifuged at ≥ 10,000*g* for 30 s. The column was washed with 700 µl Zymo DNA Wash Buffer and an additional wash with 200 µl Zymo Wash Buffer. The library was eluted with 12 µl Zymo DNA Elution Buffer. The quality and insert size of the libraries were assessed using the ds high sensitivity chip on the BioAnalyzer (Agilent).

Cluster generation and sequencing (hybrid method). The purified library was clustered on two lanes of a HiSeq4000 PE flowcell (Illumina) at a final concentration of 150 pM. The following procedure was used for clustering: libraries were diluted to 1.5 nM in H_2O and 5 µl of diluted library was added to two wells of an 8-well strip tube (Axygen). 5 µl of 0.1 N NaOH was added to the samples and incubated for 8 min at room temperature followed by the addition of 5 µl of 200 mM Tris-HCl, pH 8.0. ExAmp was performed as described in the HiSeq4000 Reagent Reference Guide. A custom Read-1 sequencing primer (IDT, standard desalting, **Supplementary Table 2**) was used at 0.1 µM final concentration in HT1 (Illumina) for clustering on an Illumina cBot. The clustered flowcell was subjected to PE sequencing (76-bp reads) on an Illumina HiSeq4000 using a custom sequencing recipe, and custom sequencing primers for barcodes and Read-2 (**Supplementary Table 2**).

Pseudo barcoding (hybrid method). In hybrid approach, the libraries from different 384-well plates with the same barcodes can be loaded onto different lanes of the flowcell. Each lane can be assigned pseudo-barcodes such that each lane represents a unique barcode space. Therefore the total barcodes can be expanded from 1 × 96 × 384 = 36,964 barcodes (for 1 lane), to 2 × 96 × 384 = 73,728 barcodes (for 2 lanes) or 3 × 96 × 384 = 110,592 barcodes (for 3 lanes).

Preparation of 150K combinatorial bead pool (single-tube method). 5′-aldehyde modified oligonucleotides 1–384 (100 µM in H_2O; **Supplementary Table 2**) were synthesized using a custom-built continuous oligonucleotide synthesizer[43,44]. 3′ Dideoxy-C (ddC) modified splint and 5′ phosphorylated oligonucleotides (100 µM in H_2O) were obtained from Integrated DNA Technologies.

Individual immobilization of each 5′-aldehyde-modified oligonucleotide (1–384) onto activated beads was performed according to a previously published method[45]. Beads from all 384 tubes were then pooled and washed three times with 0.1 × TE with 0.1% Tween 20. The pooled beads (0.5 mg × 384) were resuspended in 5 ml T7 ligase buffer (New England Biolabs), to which 7.68 ml splint oligo (100 µM) was added and mixed. 33 µl of the mixture was transferred to each tube for a total of 384 tubes. 4 µl of each individual oligonucleotide 385–768 was added to its respective tube and incubated at 80 °C for 3 min, followed by cooling on ice for 5 min. After cooling, 3 µl of T7

ligase (New England BioLabs) was added to each tube and incubated at room temperature overnight with rotation. After ligation, the tubes were placed in a heat block (SciGene) at 65 °C for 10 m to inactivate the ligase. The beads were then pooled and washed with 0.1 × TE with 0.1% Tween 20. After washing, the beads were heated at 60 °C for 2 min followed by immediate washing with 0.1 × TE with 0.1% Tween 20 to release the splint oligo. The beads were then resuspended at 5 mg/ml in 0.1 × TE buffer and stored at 4 °C.

CPT-seq on beads with 150K combinatorial bead pool (single-tube method). 50 µl of the 5 mg/ml 150K-plex bead pool was transferred with wide orifice tips (Mettler Toledo), to a strip tube (Axygen) and washed twice with 200 µl wash buffer. The beads were resuspended in 50 µl of 0.2 mg/ml BSA (New England BioLabs; NEB) in wash buffer and incubated for 10 m at room temperature on a Roto-Torque Heavy Duty Rotator. The beads were washed two times with 200 µl wash buffer and resuspended in 50 µl wash buffer. 4 µl of EZ-TSM (EZ-Tn5 with P-Hyb-ME and P-ME′ duplex) was added and mixed gently. The bead pool was incubated for 30 m at room temperature on a rotator. The beads were washed twice with 200 µl wash buffer and resuspended in 40 µl working buffer.

3 ng of freshly prepared DNA was incubated with the beads for 10 min at room temperature on a rotator. Transposition was then performed by adding 100 µl of 2× Tagment DNA buffer (Illumina) and incubating at 55 °C for 10 m. Tagmentation was stopped by adding 20 µl of 3.6 M NaCl and 20 µl of 5% SDS (Sigma Aldrich) in working buffer. The reaction was then heated at 37 °C for 10 m followed by two 200 µl washes with wash buffer. The beads were resuspended in 100 µl ELM3 (Illumina) and incubated at room temperature for 30 m on a rotator. The beads were then washed twice with 200 µl wash buffer.

PCR was performed in a single tube by removing the final wash buffer from the beads and adding 60 µl H_2O, 30 µl Nextera PCR Master Mix (NPM) (Illumina), 5 µl 10 µM P7_B15 primer, and 5 µl 10 µM P5_A14 primer. The following PCR parameters were used: an initial denaturation at 98 °C for 30 s, and 10 cycles of denaturation at 98 °C for 10 s, annealing at 63 °C for 30 s and extension at 72 °C for 1 min. The PCR was purified using 500 µl Zymo *Select-a-Size* Binding Buffer with 30 µl 95% ethanol and transferred to an IC-S Zymo *Select-a-Size* column (Zymo Research). The column was centrifuged at ≥10,000*g* for 30 s. The column was washed with 700 µl Zymo DNA Wash Buffer and an additional wash with 200 µl Zymo Wash Buffer. The library was eluted with 12 µl Zymo DNA Elution Buffer.

Cluster generation and sequencing (single-tube method). The purified library was clustered on two lanes of a HiSeq4000 PE flowcell (Illumina) at a final concentration of 150 pM. The following procedure was used for clustering: libraries were diluted to 1.5 nM in H_2O and 5 µl of diluted library was added to two wells of an 8-well strip tube (Axygen). 5 µl of 0.1 N NaOH was added to the samples and incubated for 8 min at room temperature followed by the addition of 5 µl of 200 mM Tris-HCl, pH 8.0. ExAmp was performed as described in the HiSeq4000 Reagent Reference Guide. A custom Read-1 sequencing primer was used at 0.3 µM final concentration in HT1 (Illumina) containing 40% formamide for clustering on an Illumina cBot. The clustered flowcell was subjected to PE sequencing (76-bp reads) on an Illumina HiSeq4000 using a custom sequencing recipe, and custom sequencing primers for barcodes and Read2 (**Supplementary Table 2**).

DNA capture efficiency on beads. Tagmentation was performed similar to transposition of HMW and freshly prepared NA12878 genomic DNA (gDNA), however a reduced volume was used in order to facilitate Qubit measurement. 10–30 µl of the bead pool was taken with a wide orifice pipet tip for one transposition reaction and working buffer was added to a final volume of 40 µl. With a wide-orifice pipet tip, 40 µl of HMW NA12878 gDNA (Gentra) totaling 3–50 ng was added to the beads and mixed by gentle flicking and inverting of the strip tube. The gDNA and beads were incubated at room temperature for 10 m at low speed on a rotator followed by the addition of 20 µl of 5× Tagment DNA buffer (TD, Epicentre). The sample was mixed by gentle flicking and inversion of strip tubes and incubated at 55 °C for 10 m. The 100 µl supernatant was then separated by a magnet (Invitrogen). No bead controls were made using the same quantities of input gDNA and the same ratios of tagmentation buffer and working buffer. A dsDNA HS Qubit assay (Invitrogen) was run on 10 µl of each supernatant or no bead control gDNA following manufacturer's instructions. The resulting concentrations (in ng/µl) were then multiplied by the tagmentation reaction volume (100 µl) to determine the amount of DNA not bound. The amount bound to the beads was calculated by calculating total DNA amount minus DNA not bound.

Confirmation of gene deletions by PCR. 25 ng of NA12878 gDNA was amplified with 1 U Taqurate, × LA Taq Buffer, 200 µM dNTPs, and primers (**Supplementary Table 2**) designed to span the upstream and downstream regions of the gene deletion. The following PCR conditions were used: initial denaturation at 95 °C for 4 min followed by denaturation at 95 °C for 30 s, 30 cycles of denaturation at 98 °C for 30 s, annealing at 57.5 °C for 30 s, and extension at 72 °C for 3 min, and a final extension at 72 °C for 4 min.

Phasing analysis. The details of the phasing analysis, switch error definition and calculation have been previously published[26]. The input unphased Variant Call Format (VCF) files used in this study were obtained from Amini *et al.*[26]. In short, the sequencing data were extracted from sequencing raw data and demultiplexed for each barcode. The whole-genome phasing pipeline is implemented as a Python package and processed as follows. First, the heterozygous single-nucleotide polymorphism sites (SNPs) were extracted from the unphased reference VCF file as the input, which will be later used as the markers for the linked reads to resolve their haplotype. For each barcode, paired-end reads are aligned to the reference genome and filtered with minimum mapping quality score 20 and maximum insert-size of 1.5 kb. The reads with the same index and within 15 kb proximity are assumed to be derived from the same original long DNA molecule and are merged as a linked read region. If a linked read region covers two or more input heterozygous SNPs, those SNPs are recorded as the input into H-BOP[32].

SNP/INDEL analysis and comparison. CPT-seq data were analyzed using Illumina's Isaac pipeline[46], NorthStar version 4, in default mode except that bam files were downsampled to 27× unique sequence coverage between the alignment and variant calling steps. Downsampling was performed by extracting only reads that were not duplicate-marked (samtools view -b -F 0x400 in.bam -o temp.bam, using samtools[40] v1.3) followed by extracting the fraction required to achieve 27× (sambamba[47]_v0.5.9 view —subsample = <sample-appropriate-percentage> temp.bam -o final.bam). Runs were aligned against the hg19 reference genome. The resulting gVCFs were evaluated using hap.py, version 0.2.9 (https://github.com/Illumina/hap.py), against the Platinum Genomes[35] (PG) truth set for NA12878, version 2016-1.0.

Zheng *et al.* data[11] were analyzed in two ways. In one case, the VCF file provided by Zheng *et al.* was evaluated using hap.py against the PG truth set as above. In the second case, data were run through a variation on the variant-calling pipeline described above and then evaluated as above. The bam file was converted back to fastq. Initial alignment and downsampling was performed as above. The preliminary variant calling runs exactly parallel to those described above gave disappointing results; investigation of the results indicated that poor calibration of the raw base call Qscores was at least part of the problem. As the Isaac pipeline does not have a step analogous to GATK's BQSR, the Isaac bam file was processed through BQSR using GATK version 3.1.1. The resulting recalibrated bam file was processed using an internal release of the Isaac pipeline (version 2.25.0) that had been shown to have SNP calling on par with the NorthStar v4 version and slightly improved indel performance.

Heterozygous deletions and interchromosomal translocation detection. For genomic deletion detection, the sequencing reads were demultiplexed according to their barcodes and aligned to the reference genome. Duplicate reads were removed and reads sharing the same barcode were arranged into linked read structure when the genomic distance between adjacent reads is less than 15 kb. The chromosomes were scanned with a 1 kb sliding window, in which the number of the barcodes showing any mapped reads in the scanning window, is recorded. In general, the barcode number at deletions should be lower compared to the adjacent regions. The *P*-value of the deletion can be calculated from the normal distribution estimated within 500 kb on both upstream and downstream extension excluding the center 51 kb region. Within the sliding window, the barcode can also be assigned to the haplotype using the linked read information. The number of barcodes assigned to each haplotype can then be used to calculate the *P*-value for the heterozygous deletion using the two-tailed binomial test.

doi:10.1038/nbt.3897

Interchromosomal translocation events were identified using the weighted Jaccard index sharing metric between two genomic regions. The fastq files from the CPT sequencing are demultiplexed according to their barcodes and aligned to the reference genome with the duplicates removed. The genome is scanned by 2 kb sliding window. Every 2 kb window is a 36864 vector in which each element records how many reads from a unique barcode have been found within this 2 kb window. For every 2 kb window pair (X,Y) across the genome, the weighted-Jaccard index is calculated as follows:

$$X=(x1,x2,\ldots,x36846);Y=(y1,y2,\ldots,y36846)$$

$$W\_\text{Jaccard}=\frac{\sum_i (xi+yi)\{\text{if xi}>0\text{ and yi}>0\}}{\sum_i xi+\sum_i yi}$$

Data availability. Bioproject: SUB1719920. The phasing pipeline and reference vcf can be downloaded from: https://app.box.com/folder/28851020196.

43. Lebl, M. *et al.* Automatic oligonucleotide synthesizer utilizing the concept of parallel processing. *Collect. Symp. Ser.* **12**, 264–267 (2011).
44. Kremsky, J.N. *et al.* Immobilization of DNA via oligonucleotides containing an aldehyde or carboxylic acid group at the 5′ terminus. *Nucleic Acids Res.* **15**, 2891–2909 (1987).
45. Steinberg, G., Stromsborg, K., Thomas, L., Barker, D. & Zhao, C. Strategies for covalent attachment of DNA to beads. *Biopolymers* **73**, 597–605 (2004).
46. Raczy, C. *et al.* Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041–2043 (2013).
47. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).